1.) Import the dependencies

Such as: 1) Pandas
2-) Numpy
3-) matplotlib
4-) seaborn

} Basic Data Science set

1.2-) From nltk we will import sent tokenize
word tokenize

- WordCloud
- nltk.corpus → Stopwords
- nltk.Stem → PorterStemmer

¿What is nltk?
¿What is Tokenizer?
¿What is Word Cloud?

1.3-) import the models from sklearn
1.4-) Read the Data.

} • MultinomialNB
• Random Forest Class.

2-) Understanding the Data using pandas and some Basic Data Science skills

2.1-) Delete NaN
2.2-) basic visualizations.

3-) Features Engineering.

3.2-) count the amount of words

3.3.) count the amount of characters.

3.4-) Modify the text

3.4.1-) Set it all in lowercase
3.4.2-) Remove all the URLs in the Text.
3.4.3-) Remove Punctuation.
3.4.4-) Remove stop words.
3.4.5-) Stemming

4-) Visualizations: Use matplotlib to visualize the results from the text preprocessing.

4.2) graph the most words used by each Label in the Dataset.

5.) Vectorization

5.1.) X = data["text"]

y = data["Label"]

5.2.) Use SKlearn Train_test_Split fuction on set it into 80% | 20%.

6-) Use The Machine learning models that we import at the begging of the code.

6.1-) Naive Bayes.

6.2.) RandomForest.

6.3.) Linearsvm.

6.4-) Logistic Regression.

7.) Select the Best model, The one that has the best Result.